

11/29/99
jc571 U.S. PTO

A

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Patent Application of: Luis Felipe Cabrera et al.

Title: STORAGE MANAGEMENT SYSTEM HAVING ABSTRACTED VOLUME PROVIDERS

Attorney Docket No.: 777.246US1

jc503 U.S. PTO
09/450364
11/29/99

PATENT APPLICATION TRANSMITTAL

BOX PATENT APPLICATION

Assistant Commissioner for Patents
Washington, D.C. 20231

We are transmitting herewith the following attached items and information (as indicated with an "X"):

- X Utility Patent Application under 37 CFR § 1.53(b) comprising:
X Specification (27 pgs, including claims numbered 1 through 28 and a 1 page Abstract).
X Formal Drawing(s) (4 sheets).
X Signed Combined Declaration and Power of Attorney (4 pgs).
X Check in the amount of \$982.00 to pay the filing fee.
X Assignment of the invention to Microsoft Corporation (3 pgs) and Recordation Form Cover Sheet.
X Check in the amount of \$40.00 to pay the Assignment recording fee.
X Return postcard.

The filing fee has been calculated below as follows:

	No. Filed	No. Extra	Rate	Fee
TOTAL CLAIMS	28 - 20 =	8	x 18 =	\$144.00
INDEPENDENT CLAIMS	4 - 3 =	1	x 78 =	\$78.00
] MULTIPLE DEPENDENT CLAIMS PRESENTED				\$0.00
BASIC FEE				\$760.00
TOTAL				\$982.00

Please charge any additional required fees or credit overpayment to Deposit Account No. 19-0743.

SCHWEGMAN, LUNDBERG, WOESSNER & KLUTH, P.A.
P.O. Box 2938, Minneapolis, MN 55402 (612-373-6900)

By: Rodney L. Lacy
Atty: Rodney L. Lacy
Reg. No. 41,136

Customer Number **21186**

"Express Mail" mailing label number: EL334875718US

Date of Deposit: November 29, 1999

I hereby certify that this paper or fee is being deposited with the United States Postal Service "Express Mail Post Office to Addressee" service under 37 CFR 1.10 on the date indicated above and is addressed to the Assistant Commissioner for Patents, Box Patent Application, Washington, D.C. 20231.

By: Chris Hammond

Signature: Chris Hammond

**STORAGE MANAGEMENT SYSTEM HAVING ABSTRACTED VOLUME
PROVIDERS**

“Quorum Resource Arbitrator Within A Storage Subsystem” – serial no.

5 _____ [Attorney docket 777.244US1],

“Storage Management System Having Common Volume Manager” – serial no.

_____ [Attorney docket 777.245US1],

“Volume Stacking Model” – serial no. _____ [Attorney docket 777.247US1],

10 “Volume Configuration Data Administration” – serial no. _____ [Attorney
docket 777.248US1], and

“Volume Migration Between Volume Groups” – serial no. _____ [Attorney
docket 777.249US1].

FIELD OF THE INVENTION

15 This invention relates generally to data storage, and more particularly to a storage
management system in which volume providers provide an abstracted, high-level interface for
managing storage volumes within a storage subsystem.

COPYRIGHT NOTICE/PERMISSION

20 A portion of the disclosure of this patent document contains material that is subject to
copyright protection. The copyright owner has no objection to the facsimile reproduction by
anyone of the patent document or the patent disclosure as it appears in the Patent and
Trademark Office patent file or records, but otherwise reserves all copyright rights
25 whatsoever. The following notice applies to the software and data as described below and in

BACKGROUND OF THE INVENTION

As computer systems have evolved so has the availability and configuration of data storage devices, such as magnetic or optical disks. For example, these storage devices can be connected to the computer system via a bus, or they can be connected to the computer system via a wired or wireless network. In addition, the storage devices can be separate or co-located in a single cabinet.

A storage volume is a software abstraction of underlying random access storage devices, such as magnetic or optical disks, and is the smallest self-contained unit of storage exposed by an operating system and administered by a file system. Storage volumes hide the physical topology of the storage devices and may be a fraction of a disk, a whole disk or even multiple disks that are bound into a virtually contiguous range of logical blocks. This binding may increase the fault tolerance, performance, or capacity characteristics of the underlying storage devices.

One technique for improving fault tolerance, known as mirroring or plexing a disk, uses multiple disks. When data is written to one disk the data is also written to a second disk; thus the second disk is a “mirror image” of the first disk. If one disk should fail the second disk is still available for use and has an exact copy of the information on the first disk.

Volumes are constructed from one or more extents that are contiguous storage address spaces presented by the underlying storage devices. An extent is typically characterized by the size of the address space and a starting offset for the address space from a base of the media.

Volume mapping is the process of mapping the contiguous address space presented by the volume onto the non-contiguous storage address spaces of the underlying extents. Volume mappings are either implemented on a specialized hardware controller, referred to as a hardware volume provider, or in software by a software volume provider.

5 Conventional volume providers are configured and managed with unique, vendor-specific tools. Each provider typically uses a proprietary protocol for communicating between the storage device and the driver. The volume providers often require that the administrator explicitly specify the configurations of the underlying volumes and force the administrator to make detailed, low-level configuration choices regarding the volume provider and the
10 underlying storage devices. Because of the unique protocols and management tools of conventional volume providers, it is often difficult for an administrator to manage a system having multiple storage devices. There is a need in the art for a storage management system that does not require administrators to explicitly specify volume configurations and does not require that the administrator have detailed knowledge of the volume providers and storage
15 devices.

SUMMARY OF THE INVENTION

The above-mentioned shortcomings, disadvantages and problems are addressed by the present invention. In one aspect, the invention is a storage management system in which a
20 volume provider incorporates an application programming interface (API) for receiving storage access information that characterizes high-level input/output (I/O) attributes of the storage volume such as intended access patterns, a desired level of fault handling and

optimization preferences. The volume provider and the underlying storage devices conform to the characterization as best as possible and resolve any conflicts with the communicated attributes. Furthermore, the volume providers monitor the storage volumes and may reconfigure the storage volumes in order to accomplish the communicated attributes.

5 In another aspect the invention is a method for managing one or more storage volumes of a storage subsystem. According to the invention, the method includes receiving storage access information that characterizes high-level input/output attributes of one or more of the storage volumes. The storage volumes of the storage subsystem are then configured as a function of the input/output attributes characterized by the storage access information.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 shows a diagram of the hardware and operating environment in conjunction with which embodiments of the invention can be practiced;

FIG. 2 is a block diagram illustrating a system-level overview of a storage subsystem having two nodes and a variety of storage devices;

FIG. 3 is a block diagram illustrating one embodiment of a storage management system for managing one or more storage volumes of a storage subsystem according to the invention; and

FIG. 4 is a flowchart illustrating one mode of operation of the inventive storage management system.

DETAILED DESCRIPTION OF THE INVENTION

In the following detailed description of exemplary embodiments of the invention, reference is made to the accompanying drawings that form a part hereof, and in which is shown by way of illustration specific exemplary embodiments in which the invention may be practiced. These embodiments are described in sufficient detail to enable those skilled in the art to practice the invention, and it is to be understood that other embodiments may be utilized and that changes may be made without departing from the spirit or scope of the present invention. The following detailed description is, therefore, not to be taken in a limiting sense, and the scope of the present invention is defined only by the appended claims.

The detailed description is divided into five sections. In the first section, the hardware and the operating environment in conjunction with which embodiments of the invention may be practiced are described. In the second section, a system level overview of the invention is presented. In the third section, methods for an exemplary embodiment of the invention are provided. In the fourth section, an exemplary embodiment of the invention is provided. Finally, in the fifth section, a conclusion of the detailed description is provided.

Definitions

Compromised – a status indicating that a fault tolerant volume is missing one or more disk or volume extents; for example, a mirror set with only one mirror currently available.

Configuration data - describes the mapping of physical resources to logical volumes.

Directed configuration – provider is explicitly provided with rules for choosing logical block remapping.

Disk platter – a subset of a diskpack, used for exporting or importing volumes from a diskpack.

Diskpack – a collection of logical volumes and underlying disks. A diskpack is the unit of

transitive closure for a volume.

Export – Move a disk platter and all volumes contained on that platter out of one diskpack.

Exposed – a volume is exposed to an operating system when the volume has an associated volume name (drive letter). The volume can be made available to a file system or other data store.

Free agent drive – a disk drive which is not a member of a disk pack. Free agent drives cannot contain logical volumes that are exposed.

Health – volume fault management status. A volume can be initializing, healthy, compromised, unhealthy, or rebuilding.

Healthy - containing or able to contain valid data.

Hot-spotting – temporary plexing of a volume or collection of volume extents.

Import – Move a disk platter and all volumes contained on that platter into one diskpack.

Initializing - a status indicating that a volume is rediscovering volume configuration.

LBN – logical block number.

Logical block mapping – relationship between the logical blocks exposed to the logical volume provider to those exposed by the same provider.

Logical volume provider – software which exposes logical volumes. A provider includes runtime services, configuration data, and management services.

Management service – software used to perform volume configuration, monitoring or fault handling.

Mapped volume – a simple linearly logical block mapping which concatenates volumes to expose a single larger volume.

Mirrored volume – logical volume which maintains two or more identical data copies. Also termed RAID 1.

Parity striped volume – logical volume which maintains parity check information as well as data. The exact mapping and protection scheme is vendor-specific. Includes RAID 3, 4, 5, 6.

Plexed volume – dynamic mirror volume. Plexing is used to create a copy of a volume rather

than to provide fault tolerance. The mirror is added to the volume with the intent of removal after the contents have been synchronized.

Raid - Redundant Array of Independent Disks.

Rebuilding – a status indicating that a previously compromised fault tolerant volume is resynchronizing all volume extent data.

Runtime service – software that executes on a per-IO request basis.

SCSI - Small-Computer Systems Interface.

Stacked volume – a volume that has been constructed by more than one logical block mapping operation. An example is a stripe set of mirror volumes. Stacking includes stripping, mapping, and plexing.

Striped volume – a logical block mapping which distributes contiguous logical volume extents across multiple volumes. Also termed RAID 0.

Unhealthy - a status indicating that a non-fault tolerant volume missing one or more disk or volume extents; data contained on unhealthy volumes must not be accessed.

Volume configuration stability – status to denote whether volume logical to physical mapping is undergoing change. A volume may be stable, extending, shrinking, plexing, or remapping.

Volume extent – a contiguous range of logical blocks contained on a volume. Volume extents are the smallest managed logical volume unit.

Volume status – status to denote current use of a volume by the system. A volume may be unused, hot spare, mapped, used, or unknown.

Hardware and Operating Environment

FIG. 1 is a diagram of the hardware and operating environment in conjunction with which embodiments of the invention may be practiced. The description of FIG. 1 is intended to provide a brief, general description of suitable computer hardware and a suitable computing

environment in conjunction with which the invention may be implemented. Although not required, the invention is described in the general context of computer-executable instructions, such as program modules, being executed by a computer, such as a personal computer.

Generally, program modules include routines, programs, objects, components, data structures, etc. that perform particular tasks or implement particular abstract data types.

Moreover, those skilled in the art will appreciate that the invention may be practiced with other computer system configurations, including hand-held devices, multiprocessor systems, microprocessor-based or programmable consumer electronics, network PCs, minicomputers, mainframe computers, and the like. The invention may also be practiced in distributed computing environments where tasks are performed by remote processing devices that are linked through a communications network. In a distributed computing environment, program modules may be located in both local and remote memory storage devices.

The exemplary hardware and operating environment of FIG. 1 for implementing the invention includes a general purpose computing device in the form of a computer 20, including a processing unit 21, a system memory 22, and a system bus 23 that operatively couples various system components, including the system memory 22, to the processing unit 21. There may be only one or there may be more than one processing unit 21, such that the processor of computer 20 comprises a single central-processing unit (CPU), or a plurality of processing units, commonly referred to as a parallel processing environment. The computer 20 may be a conventional computer, a distributed computer, or any other type of computer; the invention is not so limited.

The system bus 23 may be any of several types of bus structures including a memory

bus or memory controller, a peripheral bus, and a local bus using any of a variety of bus architectures. The system memory may also be referred to as simply the memory, and includes read only memory (ROM) 24 and random access memory (RAM) 25. A basic input/output system (BIOS) 26, containing the basic routines that help to transfer information
5 between elements within the computer 20, such as during start-up, is stored in ROM 24. The computer 20 further includes a hard disk drive 27 for reading from and writing to a hard disk, not shown, a magnetic disk drive 28 for reading from or writing to a removable magnetic disk 29, and an optical disk drive 30 for reading from or writing to a removable optical disk 31 such as a CD ROM or other optical media.

10 The hard disk drive 27, magnetic disk drive 28, and optical disk drive 30 are connected to the system bus 23 by a hard disk drive interface 32, a magnetic disk drive interface 33, and an optical disk drive interface 34, respectively. The drives and their associated computer-readable media provide nonvolatile storage of computer-readable instructions, data structures, program modules and other data for the computer 20. It should be appreciated by those
15 skilled in the art that any type of computer-readable media which can store data that is accessible by a computer, such as magnetic cassettes, flash memory cards, digital video disks, Bernoulli cartridges, random access memories (RAMs), read only memories (ROMs), and the like, may be used in the exemplary operating environment.

A number of program modules may be stored on the hard disk 27, magnetic disk 29,
20 optical disk 31, ROM 24, or RAM 25, including an operating system 35, one or more application programs 36, other program modules 37, and program data 38. A user may enter commands and information into the personal computer 20 through input devices such as a

keyboard 40 and pointing device 42. Other input devices (not shown) may include a microphone, joystick, game pad, satellite dish, scanner, or the like. These and other input devices are often connected to the processing unit 21 through a serial port interface 46 that is coupled to the system bus, but may be connected by other interfaces, such as a parallel port, game port, or a universal serial bus (USB). A monitor 47 or other type of display device is also connected to the system bus 23 via an interface, such as a video adapter 48. In addition to the monitor, computers typically include other peripheral output devices (not shown), such as speakers and printers.

The computer 20 may operate in a networked environment using logical connections to one or more remote computers, such as remote computer 49. These logical connections are achieved by a communication device coupled to or a part of the computer 20, the local computer; the invention is not limited to a particular type of communications device. The remote computer 49 may be another computer, a server, a router, a network PC, a client, a peer device or other common network node, and typically includes many or all of the elements described above relative to the computer 20, although only a memory storage device 50 has been illustrated in FIG. 1. The logical connections depicted in FIG. 1 include a local-area network (LAN) 51 and a wide-area network (WAN) 52. Such networking environments are commonplace in offices, enterprise-wide computer networks, intranets and the Internet.

When used in a LAN-networking environment, the computer 20 is connected to the local network 51 through a network interface or adapter 53, which is one type of communications device. When used in a WAN-networking environment, the computer 20 typically includes a modem 54, a type of communications device, or any other type of

communications device for establishing communications over the wide area network 52, such as the Internet. The modem 54, which may be internal or external, is connected to the system bus 23 via the serial port interface 46. In a networked environment, program modules depicted relative to the personal computer 20, or portions thereof, may be stored in the remote memory storage device. It is appreciated that the network connections shown are exemplary and other means of and communications devices for establishing a communications link between the computers may be used.

The hardware and operating environment in conjunction with which embodiments of the invention may be practiced has been described. The computer in conjunction with which embodiments of the invention may be practiced may be a conventional computer, a distributed computer, or any other type of computer; the invention is not so limited. Such a computer typically includes one or more processing units as its processor, and a computer-readable medium such as a memory. The computer may also include a communications device such as a network adapter or a modem, so that it is able to communicatively couple to other computers.

System Level Overview

FIG. 2 is a block diagram illustrating a system-level overview of storage subsystem 100 that includes computer 105, computer 110 and storage devices 106 that are communicatively coupled via network 120. Computers 105 and 110 represent any suitable computing system such as local computer 20 or remote computer 49 in FIG. 1. Storage devices 106 includes that includes storage device 108, storage device 112, and storage device

114 that may be any suitable storage medium such as a single internal disk, multiple external disks or even a RAID cabinet. Storage devices 106 need not be homogeneous in size, performance characteristics, or physical media.

In the embodiment illustrated by FIG. 1, storage devices 108 and 114 are directly connected to computers 105 and 110, respectively while storage device 112 is directly connected to network 120 and is a RAID storage cabinet having multiple disks. Storage devices 106 have corresponding volume providers (not shown) to map logical storage volumes onto one or more storage devices 106 of storage subsystem 100. These volume providers may be implemented in either software or hardware and may reside within the respective storage devices 108, 112 and 114 or within computers 105 and 110. As illustrated in detail below, the abstracted volume providers present a high-level interface that allows administrators to manage the volume providers and the underlying storage devices 106 without requiring that the administrator explicitly specify low-level configuration settings.

FIG. 3 is a block diagram illustrating one embodiment of an inventive storage management system 200 for managing various volume providers 203 of storage subsystem 100 (FIG.1). Storage management system 200 includes applications 210, volume providers 203 that execute in user space 218. Application 210 can be a management-type software application used by a system administrator to configure volumes or can be a user-type application that accesses the storage devices 106. As illustrated in FIG. 3, volume providers 203 consist of software volume providers 204 and hardware volume providers 206 that execute in user space 218. In other embodiments volume providers 203 execute in any combination of user space 218 and kernel space 220. Drivers 212 and 214 support run-time

services that are provided by the operating system and execute in kernel space 220. Volume providers 203 perform all management tasks such as volume configuration, performance monitoring and dynamic event handling.

Each volume provider 203 within storage management system 200 has a corresponding hardware driver 214. In addition, a software volume provider 204 interfaces to software driver 212 that controls hardware driver 214. Drivers 212 and 214 handle all communications with storage devices 106 any may implement one of a variety of protocols such as SCSI (“Small-Computer Systems Interface”), IP (“Internet Protocol”), etc. Drivers 212 and 214 support the run-time services that are provided by the operating system.

According to the invention, volume providers 203 provide applications 210 with a high-level, attribute-based view of storage devices 106. More specifically, applications 210 communicate storage access information to volume providers 203 via Application Programming Interface (“API”) 209. The storage access information allows applications 210 to communicate volume behavioral attributes instead of explicitly configuring volume providers 203. The storage access information includes input/output characteristics, data availability desires and physical mapping constraints. For example, in one embodiment the storage access information includes a desired level of fault tolerance. In another embodiment the storage access information communicated via API 209 allows applications 210 to characterize intended access patterns such as whether a volume is primarily intended for sequential reads or writes, predominantly read-only, etc. The storage access information may also include suggested configuration parameters such as request size and alignment. Finally,

via API 209, applications 210 may communicate storage access information that suggests how a volume should be optimized, such as for safety of data or for access performance.

Based on the storage access information, volume providers 203 perform all disk extent mapping, disk space allocation and mapping type selection. More specifically, volume

5 providers 203 implement sophisticated configuration rules for configuring storage devices 106 based on the storage access information received from applications 210. Thus, API 209 allows applications 210 to manage volume providers 203 without requiring that applications 210 have detailed knowledge of their underlying policies or the physical characteristics of storage devices 106. In this manner, API 209 shields applications 210 from these complex,
10 vendor-specific rules such that an administrator need not be an expert in logical volume management. In addition, API 209 allows applications 210 to manage volume providers 203 without requiring that applications 210 explicitly specify preferred volume configurations.

According to the invention, volume providers 203 are not required to configure storage device 106 so as to strictly conform to the storage access information provided by API 209.

15 To the contrary, volume providers 203 attempt to configure storage devices 106 in consideration of the storage access information communicated via management-type applications 210 as well as storage access patterns imposed by user-type applications 210. For example, a volume provider 203 could implement a single parity stripe interleave size and ignore information about additional interleave sizes. Volume providers 203 resolve
20 conflicting storage access information without limitation. In the case of a conflict, volume providers 203 attempt to complete the requested operation and report the actual configuration to applications 210.

After configuring storage devices 106, volume providers 203 monitor access patterns of applications 210 and perform dynamic reconfiguration of a volume in response to changes in the input/output (I/O) pattern. For example, based on a change in I/O pattern, volume providers 203 may move logical volume extents within storage devices 106, change

5 configuration parameters such as interleave factors, or change volume type such as substituting a parity stripe volume for a mirror set. One example is to relocate “hot-areas” of a volume, i.e. areas having a high rate of access requests, onto a set of disk that can be accessed in parallel. A second example is the identification of an area of a volume that is always accessed sequentially. Volume providers 203 may stripe such an area in order to help
10 standard read-ahead policies.

In one embodiment, however, an administrator can disable reconfiguration, effectively freezing volume configuration.

Although the volume providers 203 must conform to API 209, the invention places no restraints on how volume providers 203 communicate with the software drivers 212 or
15 hardware drivers 214 and may use any preferred protocol. In this fashion, storage management system 200 requires no changes to conventional drivers 212 and 214. Thus, vendors of storage devices 106 are free to develop their own communication protocols.

The system level overview of the operation of an exemplary embodiment of the invention has been described in this section of the detailed description. Volume providers
20 provide applications with a high-level, attribute-based view of storage devices. Furthermore, the volume providers present a high-level API that allows administrators to manage the volume providers and the underlying storage devices 106 without requiring that the

administrator explicitly specify low-level configuration settings.

Methods of an Exemplary Embodiment of the Invention

In the previous section, a system level overview of the operation of an exemplary
5 embodiment of the invention was described. In this section, the particular methods performed
by a computer executing an exemplary embodiment is described by reference to a series of
flowcharts. The methods to be performed by a computer constitutes computer programs made
up of computer-executable instructions. Describing the methods by reference to a flowchart
enables one skilled in the art to develop such programs including such instructions to carry out
10 the methods on suitable computers (the processor of the computers executing the instructions
from computer-readable media).

FIG. 4 is a flowchart 300 illustrating one exemplary method of operation of a volume
provider 203 (FIG. 3) for managing the corresponding volumes and storage devices 106 (FIG.
2) of storage subsystem 100. The volume provider 203 begins flowchart 300 with block 304.
15 In block 304, a volume provider 203 receives storage access information from applications
210 via API 209. As discussed in detail above, the storage access information characterizes
input/output (I/O) attributes of the storage volume such as intended access patterns and
optimization preferences.

In block 306 the volume provider 203 determines a configuration for storage devices
20 106 based on the storage access information including all disk extent mappings. In order to
determine the configuration, the volume provider 203 may apply industry standard rules as
well as vendor-specific rules. Industry standard rules precisely define certain configurations

such as striping with parity and mirroring. Using vendor-specific characteristics, for example, volume providers 203 can make more intelligent decisions such as which storage device 106 to use for parity in a RAID-4 configuration, how to balance the stripes across a set of disks, or which pairs of devices to use for a mirror set. In block 308, the volume provider 203

5 configures those volumes and storage devices 106 of which it has control.

In blocks 310 and 312, the volume provider 203 monitors access patterns of applications 210 and dynamically reconfigures the volumes of storage devices 106 based on previously received storage access information and any detected changes in the input/output (I/O) patterns of applications 210. For example, a volume provider 203 could reasonably
10 configure a volume as parity stripe (RAID-5) if the storage access information indicated that the volume was to be primarily read-only. The volume provider 203 could then change the parity stripe to a mirror (RAID-1) in the event that the actual load was heavily write. As another example, a volume provider could determine that fault-tolerant, single-disk volume should be concatenated or striped across two disks to distribute load. Volume providers 203
15 repeat blocks 304 through 312 upon receiving new storage access information from applications 210.

An Exemplary Embodiment of the API Provided by the Volume Providers

Listed below is one embodiment of storage access information that applications 210
20 communicate to volume providers 203 via API 209:

IsFaultTolerant - Use a fault tolerant, mirror or parity stripe volume.

FastCrashRecoveryRequired - Bound the time required for volume recovery; support for fast recovery involves use of some sort of change log enabling the provider to recover the volume without comparing the entire contents.

IsRemovable - Volume can be physically removed with minimal system disruption;

5 occupy minimal number of spindles or removable media.

OptimizeForSequentialReads - Optimize for a sequential read or sequential write pattern.

OptimizeForSequentialWrites - Optimize for a sequential write pattern. If this setting and OptimizeForSequentialReads are both false then optimize for random input/output.

10 IsMostlyReads -Optimize for read mostly pattern, for example by use of mirroring rather than parity striping.

ReconfigurationOk –Allow the extents in a volume to move within the storage device. If false, mapping of volume extents to disk extents remains fixed after configuration with the exception of actions necessary to avoid disk failures.

15 ExpectedMaximumSize - Maximum size to which the volume is likely to grow. May be greater or less than VolumeSize.

OptimalReadSize - Optimal read size in bytes with respect to the first logical block of the volume.

OptimalReadAlignment - Optimal read alignment with respect to the first logical block
20 of the volume.

OptimalWriteSize - Optimal write size in bytes with respect to the first logical block of the volume.

OptimalWriteAlignment - Optimal write alignment with respect to the first logical block of the volume.

5 MaximumNumberOfSpindles - Maximum number of disk spindles contributing to the volume. May be used to limit the number of stripe chunks in a stripe set.

InterleaveSize - Mirror or parity stripe chunk size. Number of contiguous logical blocks per disk extent prior to changing disk extents.

10 RebuildPriority - Relative priority (low:high) for rebuild. Low causes less system impact, but extends the period of exposure.

Conclusion

Various embodiments of the inventive storage management system have been described in which abstracted volume providers incorporate a generalized API by which administrators manage the volume providers and the underlying storage devices. The API allows an administrator to communicate storage access information that characterizes high-level input/output attributes of the storage devices such as intended access patterns, preferred fault handling and suggested parameters for optimization. The volume providers configure the underlying storage devices in a manner that satisfies the storage access information in light of any conflicts with the storage access information. The volume providers monitor subsequent input/output access patterns and may reconfigure the storage devices in order to

accomplish the specified high-level storage attributes. It is intended that only the claims and equivalents thereof limit this invention.

We claim:

1. A storage management system comprising a volume provider to map a logical storage volume onto one or more storage devices of a storage subsystem, wherein the volume provider presents an application programming interface (API) for receiving storage access information that characterizes behavioral attributes of the storage volume.
2. The storage management system of claim 1 wherein the attributes includes data availability desires including a desired level of fault tolerance.
3. The storage management system of claim 1 wherein the attributes include intended input/output patterns for accessing the storage volume.
4. The storage management system of claim 3 wherein the access patterns indicate whether the volume is primarily intended for sequential reads and sequential writes.
5. The storage management system of claim 1 wherein the attributes include optimization preferences.
6. The storage management system of claim 1 wherein the volume provider configures the storage volume as a function of the storage access information.

7. The storage management system of claim 1 and further including a software application executing on a computer within the storage subsystem, wherein the software application issues the storage access information to the volume provider.

5 8. The storage management system of claim 7 wherein the application is an administrative tool that issues the storage access information to the volume provider in response to input from an administrator.

10 9. The storage management system of claim 3 wherein the volume provider monitors actual access patterns and reconfigures the volume in response to changes in the actual access patterns and the intended access patterns.

15 10. The storage management system of claim 1, wherein the API conforms to a Component Object Model (COM) interface.

11. A method for managing one or more storage volumes of a storage subsystem comprising:

20 receiving via an application programming interface (API) storage access information that characterizes volume behavioral attributes of one or more storage volumes; and

configuring one or more storage volumes of a storage subsystem as a function of the storage access information.

12. The method of claim 11 wherein receiving storage access information includes receiving data availability desires including a preferred level of fault tolerance.

5 13. The method of claim 11 wherein receiving storage access information includes receiving intended access patterns.

10 14. The method of claim 13 wherein receiving the intended access patterns includes receiving whether a volume is primarily intended for sequential reads or sequential writes.

15 15. The method of claim 11 wherein receiving storage access information includes receiving configuration parameters including a request size

16 16. The method of claim 11 wherein receiving storage access information includes receiving optimization parameters.

20 17. The method of claim 16 and further including monitoring accesses of the configured storage volumes by the software application.

18. The method of claim 16 and further including reconfiguring the storage volumes based on the monitored accesses and the received storage access information.

19. The method of claim 11 wherein configuring includes resolving conflicts within the storage access information.

5 20. The method of claim 11, wherein receiving the storage access information includes receiving the storage access information via the application programming interface (API) that conforms to a Component Object Model (COM) interface.

10 21. A computer-readable medium having computer-executable instructions to cause a computer to perform a method of:
receiving via an application programming interface (API) storage access information that communicates volume behavioral attributes of one or more storage volumes; and
15 configuring one or more storage volumes of a storage subsystem as a function of the storage access information.

22. The computer-readable medium of claim 21 further including computer-executable instructions to cause the computer to further perform the method:
monitoring accesses of the configured storage volumes by the software
20 application; and
reconfiguring the storage volumes based on the accesses and the received storage access information.

23. A set of application program interfaces embodied on a computer-readable medium for execution on a computer in conjunction with a software application, comprising a first interface to receive storage access information that communicates behavioral attributes of one or more storage volumes and to configure one or more of the storage volumes of a storage subsystem as a function of the storage access information.

24. The application programming interface of claim 23 further comprising a second interface to receive data availability desires including a preferred level of fault tolerance.

25. The application programming interface of claim 24 wherein the second interface receives parameters characterizing intended access patterns for the storage volumes.

26. The application programming interface of claim 25 wherein the second interface receives parameters characterizing the intended access patterns as primarily intended for sequential reads or sequential writes.

27. The application programming interface of claim 23 and further comprising a third interface to receive optimization parameters.

28. The application programming interface of claim 23 wherein the application

programming interface (API) conforms to a Component Object Model (COM) interface.

ABSTRACT OF THE DISCLOSURE

A storage management system in which a volume provider incorporates an application programming interface (API) for receiving storage access information that characterizes high-level input/output (I/O) attributes of the storage volume such as intended access patterns, a
5 desired level of fault handling and optimization preferences. The volume provider and the underlying storage devices conform to the specified attributes as best as possible and resolve any conflicts with the communicated attributes. The volume providers monitor the storage volumes and may reconfigure the storage volumes in order to accomplish the desired attributes.

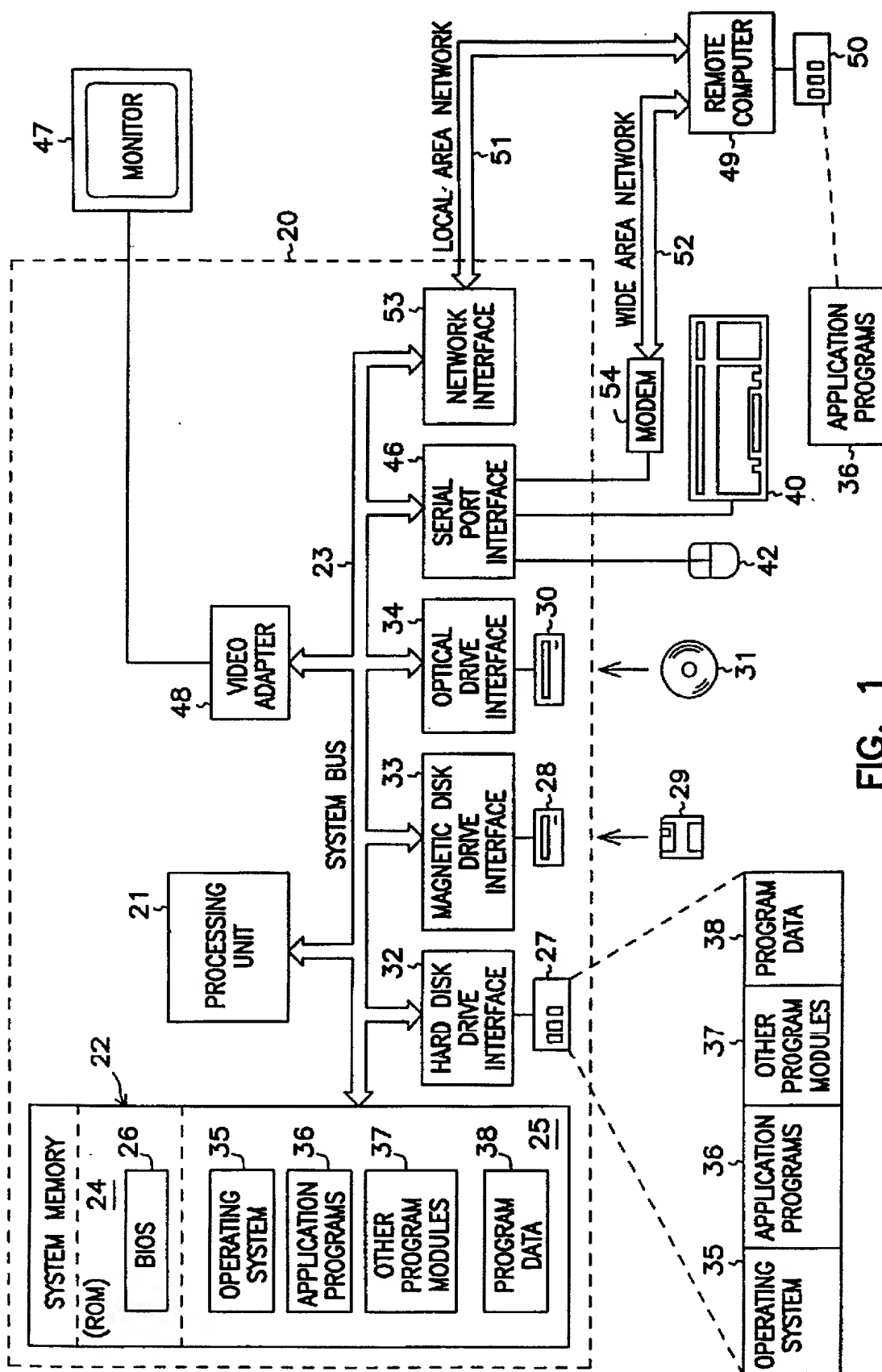
"Express Mail" mailing label number EL 334875 718 U S

Date of Deposit November 29, 1999

I hereby certify that this paper or fee is being deposited with the United States Postal Service "Express Mail Post Office to Addressee" service under 37 CFR 1.10 on the date indicated above and is addressed to the Commissioner of Patents and Trademarks, Washington, D.C. 20231

CHRIS HAMMOND
Printed Name

Chris Hammond
Signature



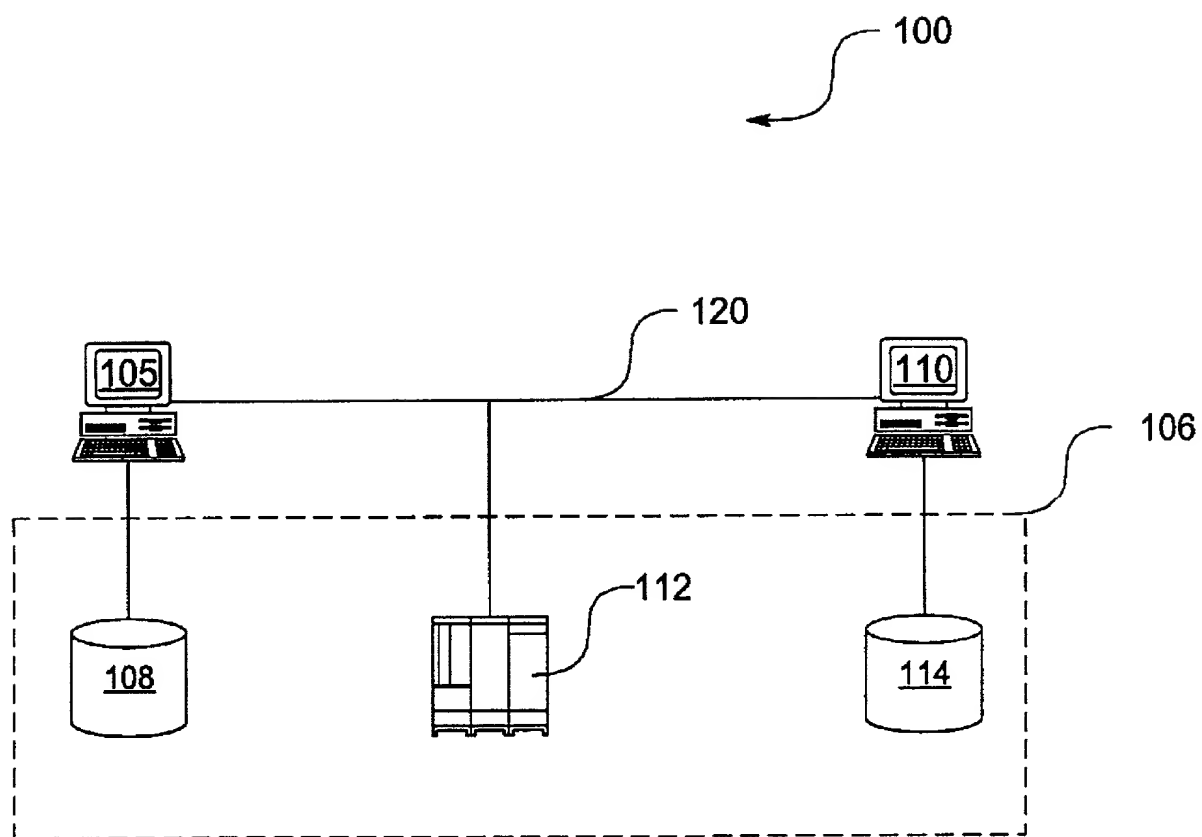


Figure 2

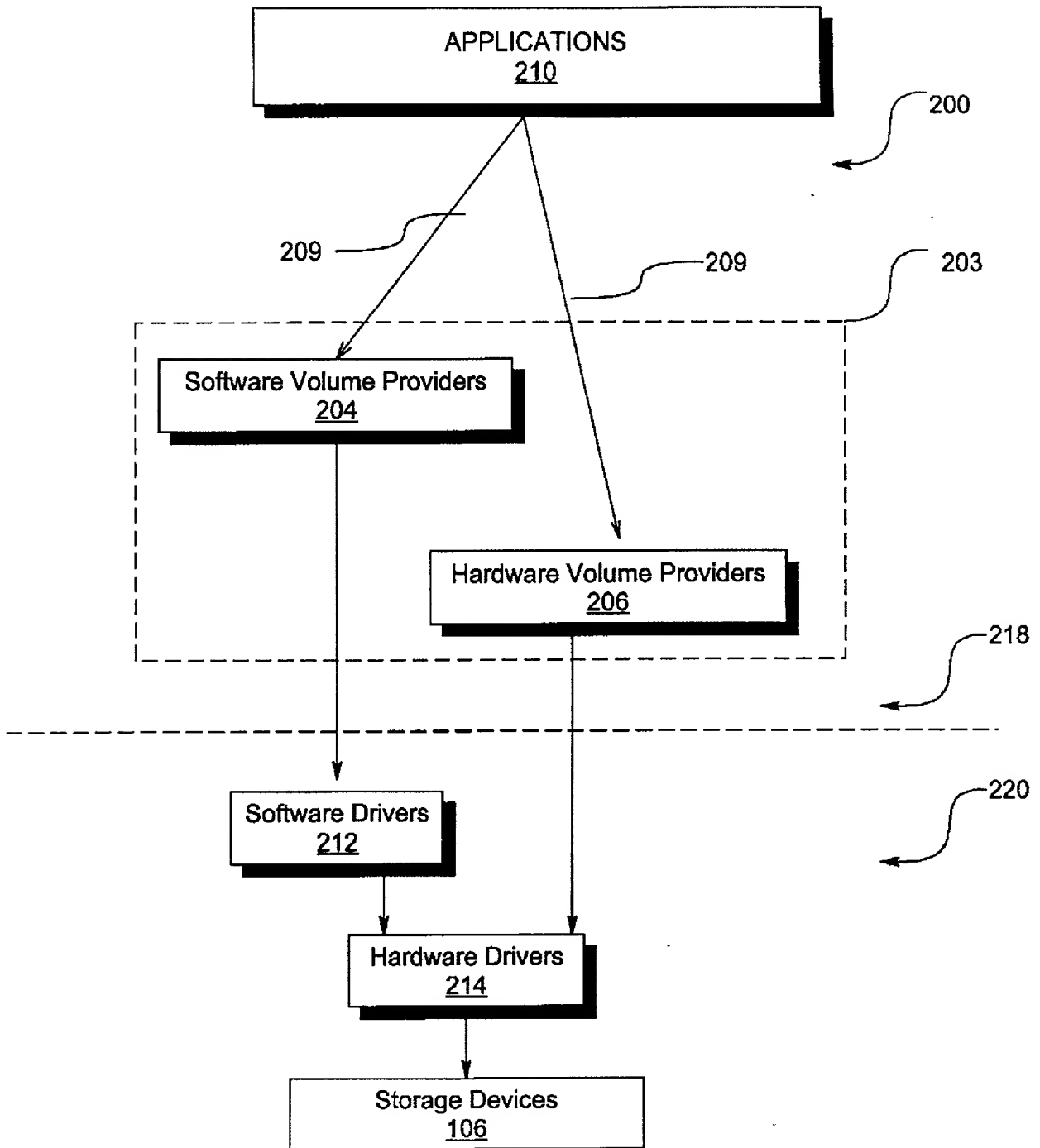


Figure 3

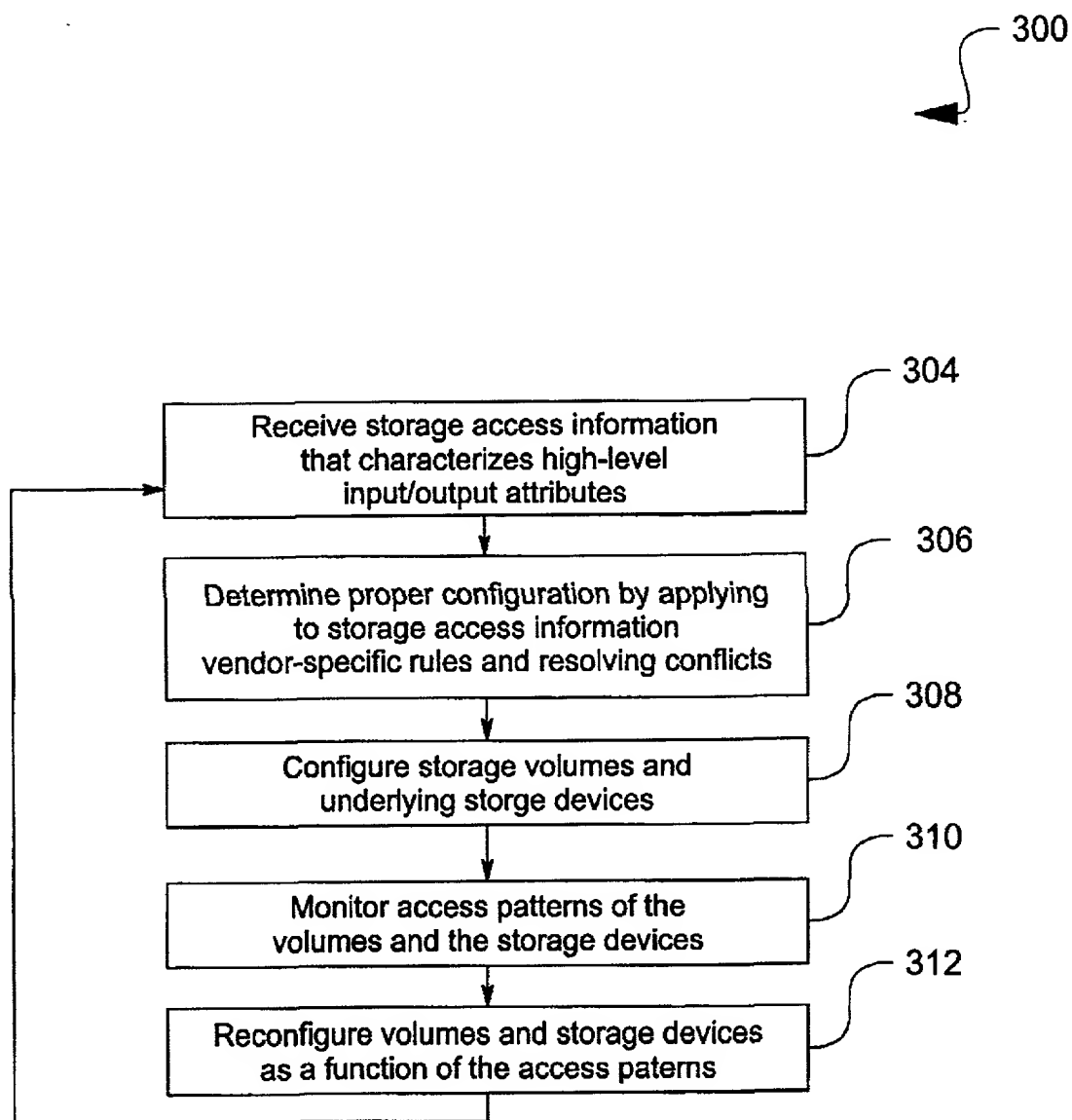


Figure 4

SCHWEGMAN, LUNDBERG, WOESSNER & KLUTH, P.A.

United States Patent Application

COMBINED DECLARATION AND POWER OF ATTORNEY

As a below named inventor I hereby declare that: my residence, post office address and citizenship are as stated below next to my name; that

I verily believe I am the original, first and joint inventor of the subject matter which is claimed and for which a patent is sought on the invention entitled: **STORAGE MANAGEMENT SYSTEM HAVING ABSTRACTED VOLUME PROVIDERS.**

The specification of which is attached hereto.

I hereby state that I have reviewed and understand the contents of the above-identified specification, including the claims, as amended by any amendment referred to above.

I acknowledge the duty to disclose information which is material to the patentability of this application in accordance with 37 C.F.R. § 1.56 (attached hereto). I also acknowledge my duty to disclose all information known to be material to patentability which became available between a filing date of a prior application and the national or PCT international filing date in the event this is a Continuation-In-Part application in accordance with 37 C.F.R. § 1.63(e).

I hereby claim foreign priority benefits under 35 U.S.C. § 119(a)-(d) or 365(b) of any foreign application(s) for patent or inventor's certificate, or 365(a) of any PCT international application which designated at least one country other than the United States of America, listed below and have also identified below any foreign application for patent or inventor's certificate having a filing date before that of the application on the basis of which priority is claimed:

No such claim for priority is being made at this time.

I hereby claim the benefit under 35 U.S.C. § 119(e) of any United States provisional application(s) listed below:

No such claim for priority is being made at this time.

I hereby claim the benefit under 35 U.S.C. § 120 or 365(c) of any United States and PCT international application(s) listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in the prior United States or PCT international application in the manner provided by the first paragraph of 35 U.S.C. § 112, I acknowledge the duty to disclose material information as defined in 37 C.F.R. § 1.56(a) which became available between the filing date of the prior application and the national or PCT international filing date of this application:

No such claim for priority is being made at this time.

I hereby appoint the following attorney(s) and/or patent agent(s) to prosecute this application and to transact all business in the Patent and Trademark Office connected herewith:

Adams, Gregory J.	Reg. No. 44,494	Forrest, Bradley A.	Reg. No. 30,837	Nama, Kash	Reg. No. 44,255
Adams, Matthew W.	Reg. No. 43,459	Harris, Robert J.	Reg. No. 37,346	Nelson, Albin J.	Reg. No. 28,650
Anglin, J. Michael	Reg. No. 24,916	Huebsch, Joseph C.	Reg. No. 42,673	Nielsen, Walter W.	Reg. No. 25,539
Arora, Suneel	Reg. No. 42,267	Jurkovich, Patti J.	Reg. No. P-44,813	Oh, Allen J.	Reg. No. 42,047
Bianchi, Timothy E.	Reg. No. 39,610	Kalis, Janal M.	Reg. No. 37,650	Padys, Danny J.	Reg. No. 35,635
Billion, Richard E.	Reg. No. 32,836	Kaufmann, John D.	Reg. No. 24,017	Parker, J. Kevin	Reg. No. 33,024
Black, David W.	Reg. No. 42,331	Klima-Silberg, Catherine I.	Reg. No. 40,052	Peacock, Gregg A.	Reg. No. P-45,001
Brennan, Leoniede M.	Reg. No. 35,832	Kluth, Daniel J.	Reg. No. 32,146	Perdok, Monique M.	Reg. No. 42,989
Brennan, Thomas F.	Reg. No. 35,075	Lacy, Rodney L.	Reg. No. 41,136	Polglaze, Daniel J.	Reg. No. 39,801
Brooks, Edward J., III	Reg. No. 40,925	Leffert, Thomas W.	Reg. No. 40,697	Prout, William F.	Reg. No. 33,995
Chu, Dinh C.P.	Reg. No. 41,676	Lemaire, Charles A.	Reg. No. 36,198	Sako, Katie E.	Reg. No. 32,628
Clark, Barbara J.	Reg. No. 38,107	Litman, Mark A.	Reg. No. 26,390	Schwegman, Micheal L.	Reg. No. 25,816
Crouse, Daniel D.	Reg. No. 32,022	Lundberg, Steven W.	Reg. No. 30,568	Steffert, Kent J.	Reg. No. 41,312
Dahl, John M.	Reg. No. P-44,639	Mack, Lisa K.	Reg. No. 42,825	Slifer, Russell D.	Reg. No. 39,838
Drake, Eduardo E.	Reg. No. 40,594	Maki, Peter C.	Reg. No. 42,832	Steffey, Charles E.	Reg. No. 25,179
Elisevva, Maria M.	Reg. No. 43,328	Malen, Peter L.	Reg. No. P-44,894	Terry, Kathleen R.	Reg. No. 31,884
Embretson, Janet E.	Reg. No. 39,665	Mates, Robert E.	Reg. No. 35,271	Viksnins, Ann S.	Reg. No. 37,748
Fogg, David N.	Reg. No. 35,138	McCrackin, Ann M.	Reg. No. 42,858	Woessner, Warren D.	Reg. No. 30,440
Fordenbacher, Paul J.	Reg. No. 42,546				

I hereby authorize them to act and rely on instructions from and communicate directly with the person/assignee/attorney/firm/organization/who/which first sends/sent this case to them and by whom/which I hereby declare that I have consented after full disclosure to be represented unless/until I instruct Schwegman, Lundberg, Woessner & Kluth, P.A. to the contrary.

Please direct all correspondence in this case to **Schwegman, Lundberg, Woessner & Kluth, P.A.** at the address indicated below:
P.O. Box 2938, Minneapolis, MN 55402
Telephone No. (612)373-6900

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

Full Name of joint inventor number 1: **Luis Felipe Cabrera**

Citizenship: **United States of America**
Post Office Address: **2009 Killarney Way SE.
Bellevue, WA 98004**

Residence: **Bellevue, WA**

Signature: _____

Luis Felipe Cabrera

Date: _____

10/18/99

Full Name of joint inventor number 2: **Norbert P. Kusters**

Citizenship: **Canada**
Post Office Address: **19310 NE 129th Way
Woodinville, WA 98072**

Residence: **Woodinville, WA**

Signature: _____

Norbert P. Kusters

Date: _____

10/15/99

X Additional inventors are being named on separately numbered sheets, attached hereto.

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

Full Name of joint inventor number 3 : **Catharine van Ingen**

Citizenship: **United States of America**

Residence: **Berkeley, CA**

Post Office Address: **3031 Fulton Street
Berkeley, CA 94705**

Signature: **Catharine van Ingen**
Catharine van Ingen

Date: **15 OCT 99**

Full Name of inventor:

Citizenship:

Residence:

Post Office Address:

Signature: _____

Date: _____

Full Name of inventor:

Citizenship:

Residence:

Post Office Address:

Signature: _____

Date: _____

Full Name of inventor:

Citizenship:

Residence:

Post Office Address:

Signature: _____

Date: _____

§ 1.56 Duty to disclose information material to patentability.

(a) A patent by its very nature is affected with a public interest. The public interest is best served, and the most effective patent examination occurs when, at the time an application is being examined, the Office is aware of and evaluates the teachings of all information material to patentability. Each individual associated with the filing and prosecution of a patent application has a duty of candor and good faith in dealing with the Office, which includes a duty to disclose to the Office all information known to that individual to be material to patentability as defined in this section. The duty to disclose information exists with respect to each pending claim until the claim is canceled or withdrawn from consideration, or the application becomes abandoned. Information material to the patentability of a claim that is canceled or withdrawn from consideration need not be submitted if the information is not material to the patentability of any claim remaining under consideration in the application. There is no duty to submit information which is not material to the patentability of any existing claim. The duty to disclose all information known to be material to patentability is deemed to be satisfied if all information known to be material to patentability of any claim issued in a patent was cited by the Office or submitted to the Office in the manner prescribed by §§ 1.97(b)-(d) and 1.98. However, no patent will be granted on an application in connection with which fraud on the Office was practiced or attempted or the duty of disclosure was violated through bad faith or intentional misconduct. The Office encourages applicants to carefully examine:

- (1) prior art cited in search reports of a foreign patent office in a counterpart application, and
- (2) the closest information over which individuals associated with the filing or prosecution of a patent application believe any pending claim patentably defines, to make sure that any material information contained therein is disclosed to the Office.

(b) Under this section, information is material to patentability when it is not cumulative to information already of record or being made of record in the application, and

- (1) It establishes, by itself or in combination with other information, a prima facie case of unpatentability of a claim; or
- (2) It refutes, or is inconsistent with, a position the applicant takes in:
 - (i) Opposing an argument of unpatentability relied on by the Office, or
 - (ii) Asserting an argument of patentability.

A prima facie case of unpatentability is established when the information compels a conclusion that a claim is unpatentable under the preponderance of evidence, burden-of-proof standard, giving each term in the claim its broadest reasonable construction consistent with the specification, and before any consideration is given to evidence which may be submitted in an attempt to establish a contrary conclusion of patentability.

(c) Individuals associated with the filing or prosecution of a patent application within the meaning of this section are:

- (1) Each inventor named in the application;
- (2) Each attorney or agent who prepares or prosecutes the application; and
- (3) Every other person who is substantively involved in the preparation or prosecution of the application and who is associated with the inventor, with the assignee or with anyone to whom there is an obligation to assign the application.

(d) Individuals other than the attorney, agent or inventor may comply with this section by disclosing information to the attorney, agent, or inventor.